PREDICT WHETHER A VARIANT WILL HAVE CONFLICTING CLINICAL CLSSIFICATION

ABSTRACT

Variant classification is sequence-based testing of the germ-line DNA to determine whether an individual carries a variation in a gene that is likely to cause a problem in normal gene function. The classification of the variants is condensed into three categories i.e benign or likely benign, Variant of Uncertain Significance(VUS), and likely pathogenic or pathogenic. When different labs submit the same category for the same variant, it is called a consistent classification, whereas when different labs submit any two of the three categories for the same variant it is considered as conflicting classification.

Although there exist international guidelines for classifying genetic variants, analysis of different laboratories cause conflicting clinical classification leading to flawed medical management. Our objective is to apply machine learning to identify the conflicting variant by comparing different lab results and predicting whether a genetic variant will have a conflicting clinical classification in future analyses.

The dataset that is used for this analysis is collected from ClinVar and Kaggle(https://www.kaggle.com/kevinarvai/clinvar-conflicting) and has a binary representation in which 0 denotes consistent classification whereas 1 denotes conflicting classification.

To predict whether a variant will have conflicting clinical classification we can implement a supervised machine learning method i.e. classification technique in predictive analysis . By applying 'classification technique', we can categorize whether the variant is consistent or conflicting which will help in future analyses . For this analysis we can use various tools such as Python, R, Weka in which we can run different classification algorithms to obtain the accurate result.